Journal of the Faculty of Engineering and Architecture of Gazi University 40:1 (2025) 413-427

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Prediction of COVID-19 disease severity using synthetic data oversampling and machine learning methods on data at first hospitalization

Kübra Köksal[1]*, Buket Doğan[1], Zehra Aysun Altıkardeş[2,3]
[1]Department of Computer Engineering, Faculty of Technology, Marmara University, 34722, İstanbul, Türkiye
[2]Department of Computer Technologies, Vocational School of Technical Sciences, Marmara University, 34722, İstanbul, Türkiye
[3]Marmara University Hypertension and Atherosclerosis Research Center (HIPAM), Maltepe, İstanbul, Türkiye

Highlights:

- Predicting WHO-CPS oriented disease severity and/or progression in hospitalized COVID-19 patients
- Detecting COVID-19 respiratory and intensive care with a high accuracy performance with fewer features
- Determining most appropriate laboratory features related intensive care and oxygen requirements

Keywords:

- Covid-19
- Machine Learning
- Laboratory data
- Feature Selection
- SMOTE

Article Info:

Research Article
Received: 23.08.2023
Accepted:19.03.2024

DOI:

10.17341/gazimmfd.1348341

Acknowledgement:

Prof. Dr. Director of Marmara University Hypertension and Atherosclerosis Training, Application and Research Center (HIPAM) for their support in data collection and research process design in the study. Dr. We would like to thank Ali Serdar Fak and his researchers.

Correspondence:

Author: Kübra Koksal
e-mail: koksalkubra0@gmail.com
phone: +90 543 209 9001

Graphical/Tabular Abstract

In this study, first of all, laboratory and demographic data are collected from patients who apply to the hospital with the diagnosis of Covid-19. After deleting the missing and inconsistent data, the pre-processing phase is completed by filling in the missing data according to the average and day intervals. Using all the features on the obtained data set and using the features selected by the Random Feature Selection algorithm, KNN, Random Forest, Bagging and Decision Tree models are created and the patient's oxygen and intensive care requirements are estimated using the WHO target value. Finally, by applying SMOTE to the data set obtained with the selected features, all results are compared in terms of accuracy and F1-Score values. The graphical summary of the study, which basically consists of 4 stages, is given in Figure A.

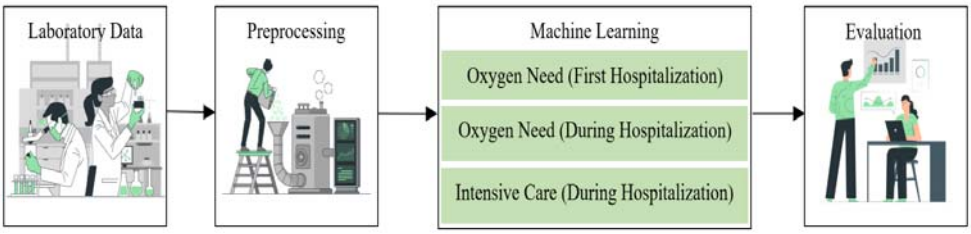


Figure A. Covid-19 intensive care and oxygen requirement prediction using machine learning

Purpose:
To develop and test machine learning models that help predict WHO-oriented disease severity by using the laboratory and demographic characteristics of patients infected with COVID-19 at the hospital admission stage.

Theory and Methods:
Preprocessing is performed on laboratory and demographic data collected from patients who applied to the hospital with the diagnosis of Covid-19. Then, KNN, Random Forest, Bagging and Decision Tree machine learning algorithms are used on the data set to predict the oxygen requirement at the first hospitalization (Analysis-1) and the oxygen requirements (Analysis-2) and intensive care (Analysis-3) during the hospitalization period. In addition, the results obtained by applying Random Feature Selection and SMOTE in the data set were compared.

Results:
The random forest algorithm is given the best results for all three analyses. Analysis-1 91.67% with 16 features, Analysis-2 91.96% accuracy with 18 features, and analysis-3 91.96% accuracy with 12 features is reached. MDW and Troponin T-hs were the most relevant features selected in common in all three analyses. When the F1-Score values of the minority classes were analyzed after applying SMOTE, an increase of 6% was observed in Analysis-1, a 24% increase in Analysis-2 and a 21% increase in Analysis-3.

Conclusion:
The Random Forest algorithm has shown a high performance in the Covid-19 intensive care and oxygen therapy prediction area. In addition, by using fewer features, the accuracy values reached when all features are used can be obtained.

Journal of the Faculty of Engineering and Architecture of Gazi University 40:1 (2025) 413-427

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

İlk yatıştaki veriler üzerinde yapay veri çoğaltma ve makine öğrenmesi yöntemleri kullanılarak COVID-19 hastalık şiddetinin tahmini

Kübra Köksal[1]*, Buket Doğan[1], Zehra Aysun Altıkardeş[2,3]

[1]Marmara Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği, 34722, Kadıköy, İstanbul, Türkiye
[2]Marmara Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, 34722, Kadıköy, İstanbul, Türkiye
[3]Marmara Üniversitesi Hipertansiyon ve Ateroskleroz Eğitim, Uygulama ve Araştırma Merkezi (HİPAM), Maltepe, İstanbul, Türkiye

ÖNEÇIKANLAR

- COVID-19 hastalık şiddeti tahmini
- En ilişkili özelliklerin belirlenmesi
- Yapay veri çoğaltmanın sınıflandırma performansına etkisinin değerlendirilmesi

Makale Bilgileri

Araştırma Makalesi
Geliş: 23.08.2023
Kabul: 19.03.2024

DOI:
10.17341/gazimmfd.1348341

Anahtar Kelimeler:
COVID-19,
makine öğrenmesi,
laboratuvar verisi,
prognoz

ÖZ

2019 Aralık ayında Çin'in Wuhan kentinde ortaya çıkan ve 11 Mart 2020'de Dünya Sağlık Örgütü tarafından pandemi olarak ilan edilen COVID-19, dünya genelinde hızla yayılarak başta sağlık sektörü olmak üzere insan hayatını olumsuz etkileyecek bir sürecin başlamasına neden olmuştur. Bu çalışmada da Marmara Üniversitesi Hastanesine başvuran COVID-19 enfekte hastaların hastaneye kabul aşamasındaki laboratuvar ve demografik özellikleri kullanılarak WHO (World Health Organization) odaklı hastalık şiddeti tahmin modelinin geliştirilmesi amaçlanmıştır. Veri seti üzerinde oksijen ve yoğun bakım ihtiyacı sonlanım durumları ile laboratuvar sonuçları arasındaki ilişki K-En yakın komşu, Torbalama, Rastgele Orman ve Karar Ağacı makine öğrenmesi yöntemleri kullanılarak analiz edilmiştir. Veri setindeki dengesiz sınıf dağılımı SMOTE metodu kullanılarak dengelenmiştir ve veri çoğaltmanın sınıflandırma performansına etkisi değerlendirilmiştir. SMOTE uygulanmayan veri setinde hastanın ilk yatış aşamasındaki oksijen ihtiyacı (Analiz – 1) 16 özellik ile %91,67, yatış sırasındaki oksijen ihtiyacı (Analiz – 2) 18 özellik ile %91,96 ve yatış sırasındaki yoğun bakım ihtiyacı (Analiz – 3) 12 özellik ile %92,17 doğruluk değeri ile tahmin edilmiştir. SMOTE veri çoğaltma işleminden sonra F1-Skor değerlerinde sırasıyla %6'lık, %24'lük ve %21'lik bir artış gözlenmiştir. Bu çalışma; COVID-19'un tanısı, izlenmesi ve klinik yönetimi için hangi laboratuvar testleri gerektiği konusunda hasta verilerine dayalı makine öğrenmesi yöntemlerine ait sonuçları ile bu alana önemli katkılar sağlamaktadır.

Prediction of COVID-19 disease severity using synthetic data oversampling and machine learning methods on data at first hospitalization

HIGHLIGHTS

- COVID-19 disease severity prediction
- Identifying the most relevant attributes
- Evaluation of the effect of synthetic data augmentation on classification performance

Article Info

Research Article
Received: 23.08.2023
Accepted:19.03.2024

DOI:
10.17341/gazimmfd.1348341

Keywords:
COVID-19,
machine learning,
laboratory data,
prognosis

ABSTRACT

COVID-19, originating in Wuhan, China, in December 2019 declared a pandemic by the World Health Organization on March 11, 2020, rapidly spread worldwide, significantly impacting human life and the health sector. This study aims to develop a WHO (World Health Organization) oriented disease severity prediction model using laboratory and demographic data from COVID-19 patients upon admission to Marmara University Hospital. The relationship between oxygen and intensive care needs with laboratory results on the data set was analyzed using K-nearest neighbor, Bagging, Random Forest and Decision Tree machine learning methods. The dataset's unbalanced class distribution was balanced using the SMOTE method, and the impact of data multiplication on classification performance was evaluated. In the data set without SMOTE, the patient's oxygen requirement during the first hospitalization was estimated with 16 features at 91.67% accuracy, the oxygen requirement at hospitalization with 18 features at 91.96%, and the intensive care need at hospitalization with 12 features at 92.17% accuracy. After SMOTE data multiplication, an increase of 6%, 24% and 21% was observed in F1-Score values, respectively. This study significantly contributes to the field by utilizing machine learning methods on patient data, essential for COVID-19 diagnosis, monitoring, and clinical management through required laboratory tests.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors :*koksalkubra0@gmail.com, buketb@marmara.edu.tr, aaltikardes@marmara.edu.tr / Tel: +90 543 209 9001

1. Giriş (Introduction)

Şiddetli akut solunum sendromu koronavirüs 2 (SARS-CoV-2) virüsü ile ilişkili olan COVID-19, dünya çapında bir pandemi haline gelmeden önce ilk olarak 2019 yılında Çin'in Wuhan kentinde ortaya çıkmıştır. COVID-19'un ölümcüllük oranı %2 değeri ile koronavirüs ailesinden gelen diğer virüs çeşitlerine (Sars-Cov (%9,5), Mers-Cov (%35)) göre daha düşük [1] olmasına rağmen Aralık 2019'dan itibaren hızlı bir şekilde yayılarak tüm dünyayı etkisi altına almıştır. [2]. 2020 yılının Mart ayında Dünya Sağlık Örgütü tarafından pandemi olarak nitelendirilmiştir [3]. COVID-19 pandemisinin ortaya çıktığı tarihten 2023 yılına kadar, dünya çapında 651 milyondan fazla onaylanmış vaka, altı milyondan fazla ölüm bildirilmiştir ve yaklaşık 13 milyar doz aşı yapılmıştır [4]. Bu eşi görülmemiş ve beklenmedik vaka artışı nedeniyle dünyanın dört bir yanındaki sağlık uzmanları ve sağlık tesisleri zorlu bir süreç geçirmek zorunda kalmıştır [5]. COVID-19'un heterojen klinik özellikler ile ortaya çıkabilmesi ve çoklu organ hasarına neden olmasından dolayı bu zorlu süreç boyunca etkili triyaj uygulanması, zamanında risk sınıflandırılmasının yapılması büyük bir önem teşkil etmekte ve her bir hasta için hayati önem taşımaktadır [6, 7]. Pandemi boyunca çoğunlukla ileri yaştaki hastalar başta olmak üzere COVID19 şüphesi bulunan birçok insanın teşhis ve tedavi işlemleri için sağlık kuruluşlarına başvurması gerekmiştir. Artan vaka ve hasta sayısı ile birlikte sağlık sektörü üzerinde ciddi bir yük oluşmuştur. Bu da bazı ülkelerde teşhis, tedavi süreçlerinin uzamasına ve sağlık sektörünün görevini yerine getiremeyecek hale gelmesine sebep olmuştur. Bu zorlu süreçte sağlık, bilgisayar bilimleri gibi alanlar çeşitli analizler yaparak sağlık sektörü üzerinde yükü hafifletmeyi ve hastaların tedavi süreçlerinde fayda sağlamayı hedeflemiştir. Sıklıkla kullanılan görüntü, laboratuvar, ses gibi klinik veriler kullanılarak yapılan analizlerin yanında COVID-19'un sosyal, duygusal, mekânsal gibi bir çok açıdan da insan hayatına etkisi analiz edilmiştir. Çılgın vd.nin [8] COVID-19 aşılarına yönelik duygu analizi ve Sönmez vd.nin [9] hastanelerdeki mekan için planlama ve mekan havalandırma üzerine elde ettikleri sonuçlar COVID-19'un sosyal ve mekânsal alanlarda yapılan analizlere örnek olarak verilebilir. Bu çalışmada ise sıklıkla kullanılan laboratuvar klinik verileri kullanılmıştır. Çeşitli klinik veriler ile COVID-19 arasında kurulacak ilişki bilgisi kullanılarak COVID-19'un erken teşhisinin gerçekleştirilmesi veya solunum desteğine ihtiyaç duyacak veya daha kötüleşebilecek hastalar için risk sınıflandırılması yapılmasının sağlanması önemli bir konudur. Teşhis süreçlerini hızlandırabilecek veya çeşitli analizler ile birlikte solunum cihazı, yoğun bakım üniteleri gibi tıbbi kaynakların en doğru şekilde kullanılmasını sağlayacak her türlü bilimsel çalışma bu süreçte büyük önem arz etmektedir. Bu konuda COVID-19 hastalarının klinik verileri üzerinde makine öğrenmesi yöntemlerinin kullanımını içeren birçok farklı çalışma gerçekleştirilmiştir. Bu çalışmalardan COVID-19'un tespiti, solunum ihtiyacı ve ölüm durumunun makine öğrenmesi ile öngörülmesine yönelik literatürdeki önemli örneklere bu başlık altında yer verilmektedir.

Banerjee vd. [10] laboratuvar sonuçları ve demografik verileri üzerinde lojistik regresyon, rastgele orman, yapay sinir ağları ve glmnet makine öğrenmesi yöntemlerini kullanarak COVID-19 hastalığını teşhis etmeyi hedeflenmiştir. Tam kan sayımını kullanan tanı prosedüründe yapay sinir ağları %95, glmnet %94 ve rastgele orman %94 AUC değerine ulaşmıştır. Hastaneye kabul edilmeyen hastalarda tanı sürecinde yapay sinir ağları %82, glmnet %84 ve rastgele orman %86 AUC değerlerine ulaşmıştır. COVID-19 hastalığını teşhis etmek için Mondal vd. [11] tarafından gerçekleştirilen çalışmada RT-PCR ve ek laboratuvar testleri ile Destek vektör makineleri, K-en yakın komşu, XGBoost, çok katmanlı algılayıcı (MLP), lojistik regresyon, karar ağaçları, rastgele orman, topluluk yöntemleri, lineer regresyon ve polinomsal regresyon yöntemleri kullanılmıştır. Brezilya'daki İsrail Hastanesi Albert Einstein tarafından sağlanan 5644 veri üzerinde analiz işlemi gerçekleştirilmiştir. MLP, lojistik regresyon ve XGBoost %91 doğrulukla en iyi sınıflandırma modelleri olarak seçilmiştir. Akarsu [12] ise 510 hastaya ait hematokrit, hemoglobin, trombosit gibi 15 farklı kan testi bilgisinden oluşan veri seti üzerinde yapay zeka ve makine öğrenmesi algoritmalarını kullanarak COVID-19'u tespit etmeyi hedeflemiştir. Oluşturulan model COVID-19 sınıflandırma alanında %89,6 ile yüksek bir sınıflandırma doğruluğuna ulaşmıştır.

COVID-19 hastalarının entübasyon ve solunum durumlarını inceleyen çalışmalar incelendiğinde Arvind vd.nin [13] hayati değerleri, laboratuvar ve demografik verileri kullanarak COVID-19 pozitif olan veya şüpheli hastaların gelecekteki entübasyon oranını tahmin etmek için bir Rastgele Orman modeli geliştirildiği görülmektedir. Oluşturulan modelin performansı ROX indeksi ile karşılaştırılmıştır. Beş hastaneye başvuran toplam 4087 hasta üzerinde analiz işlemleri yapılmıştır. Kaplan-Meier analizi sonucunda hastaneye yatış sırasında entübasyon ihtiyacının daha yüksek olduğu görülmüştür. Kullanılan makine öğrenmesi algoritması 0,83 AUC ve 0,32 AUPRC değeri ile entübasyon riski açısından ROX endeksinden önemli ölçüde daha iyi performans göstermiştir. Burdick vd. ise [14] çeşitli hayati değerleri, laboratuvar verileri üzerinde XGBoost makine öğrenmesi yöntemini kullanarak solunum dekompansasyonu tahmin etmiştir. Önerilen algoritma, 0,78 tanısal olasılık oranına sahip ventilasyonu tahmin etmek için kullanılan bir erken uyarı sistemi olan Modifiye Erken Uyarı Puanına (MEWS) göre 0,90 değeri ile daha yüksek bir performans ve duyarlılığa ulaşmıştır. Sonuçlar COVID-19 hastalarının 24 saat içinde mekanik ventilasyon ihtiyacının doğru bir şekilde tahmin edildiğini göstermiştir. Di Castelnuovo vd. ise [15] demografik değerleri ve hastanın tıbbi öyküsü (kronik akciğer, diyabet, hipertansiyon vb.) ile ilgili verileri kullanılarak COVID-19 hastalarında ölüm ihtimalini tahmin etmiştir ve en belirleyici faktörleri belirlemiştir. Rastgele orman modeli %95,2 hassasiyet, %30,8 özgüllük, %83,4 sınıflandırma doğruluğu ve %90,4 F1 değeri ile güçlü bir tahmin başarısı elde etmiştir. Ayrıca en tanımlayıcı özelliklerin bozulmuş böbrek fonksiyonu, yüksek C reaktif proteini ve ileri yaş olduğu bulunmuştur. Obezite, tütün kullanımı, kardiyovasküler hastalık ve buna bağlı komorbiditelerle ilgili bir ilişki görülmemiştir. Özellikle 85 yaşındaki hastaların 18-64 yaş arasındaki hastalara göre 8 kat daha yüksek mortaliteye sahip olduğu sonucuna ulaşılmıştır. Bu sonuçlar çok değişkenli Cox hayatta kalma analizi kullanılarak doğrulanmıştır.

COVID- 19 hastalığının seviyelerinin belirlenmesi ile ilgili çalışmalardan Huyut'un [16] gerçekleştirdiği çalışmada, Mart-Eylül 2021 COVID-19 tanısı ile hastaneye yatırılan ağır ve hafif enfekte hastalardan oluşan bir veri seti kullanılmıştır. Bu çalışmada COVID-19'un prognozunu etkileyen rutin kan değerleri ve demografik veriler kullanılarak kabul sırasında ciddi-hafif enfekte hastaları çeşitli sınıflandırma yöntemlerini kullanılarak tahmin edilmeye çalışılmıştır. Çalışmada 28 rutin kan değer parametresi ve yaş değişkeni en etkili özellikler olarak bulunmuştur. Hasta gruplarını en yüksek AUC değeri ile tahmin eden modeller sırasıyla şu şekildedir: yerel ağırlıklı öğrenme (LWL) %0,95, K-Star (K*) %0,91, naive bayes (NB) 0,85 % ve K-en yakın komşu (KNN) %0,75.

Gözde vd. 'nin [17] çalışmasında ise beş farklı risk faktör ile COVID-19 ölüm hızı arasındaki ilişki DEMATEL yöntemi kullanılarak analiz edilmiştir. Sonuçlar komorbiditenin COVID-19 ölüm oranı üzerindeki etkisinin en fazla olduğunu göstermiştir. Komorbiditeden sonra bağışıklık sistemi ve yaş da ölüm oranında önemli bir etki derecesine sahiptir. Bu faktörlerin yanında sigara kullanımının

nispeten daha düşük bir etkiye sahip olduğu ve cinsiyetin en az etkiye sahip kriter olduğu sonucuna varılmıştır.
Literatürdeki örneklerden yola çıkarak, farklı hayati değerler, laboratuvar ve demografik veriler kullanılarak COVID-19'un yüksek doğrulukta tespit edilebildiği, solunum ihtiyacı, solunum dekompansasyonu ve ölüm oranı gibi oranların başarı ile tahmin edilebildiği görülmüştür. Ayrıca daha uygun ve çabuk ölçülebilen rutin kan değerlerinin birçok viral hastalığın tanı ve prognozunda kullanıldığı bilinmektedir [18-20]. Çalışmalarda Rastgele Orman (RF), Karar Ağacı (DT), Destek Vektör Makineleri (SVM), XGBoost algoritmaları yüksek performanslı modeller oluşturulurken kullanılan makine öğrenmesi yöntemleridir. Kan gazı ölçümleri, yaşamsal belirtiler, ileri yaş, minimum oksijen satürasyonu, kan grubu gibi klinik parametreler bu çalışmalar sonucunda bulunan önemli niteliklere örnek olarak verilebilir.

Bu çalışma kapsamında hastaneye yatırılan COVID-19 ile enfekte olmuş hastaların hastaneye kabulde alınan temel klinik ve laboratuvar özellikleri kullanılarak WHO (World Health Organization) odaklı hastalık şiddetini ve/veya ilerlemesini tahmin etmeye yardımcı olan makine öğrenmesi modellerinin geliştirilmesi, test edilmesi ve elde edilen anlamlı bilgiler sayesinde pandemi döneminde oksijen cihazı ve yoğun bakım ünitesi gibi tıbbi kaynakların daha yararlı ve etkili bir şekilde kullanılması, 24 saat içerisinde alınarak analiz edilen laboratuvar sonuçları sayesinde etkili triyaj ve zamanında risk sınıflandırılması işlemlerinin gerçekleştirilmesi konularında katkı sağlanması hedeflenmektedir.

Bu çalışma, özellikle gelişmekte olan ülkelerde makine öğrenmesi yöntemleri ile Covid-19 hastalık şiddetinin tahmini için gerekli önemli laboratuvar testlerinin öngörülmesi, erken ve uygun sağlık yönetiminin sağlanması, optimal kaynak tahsisini teşvik etmeye yardımcı olabilecek hasta laboratuvar analizleri tanımlamayı amaçlayan, hastalığın taranması ve izlenmesi için kritik destek sağlayabilecek, hastane ortamında sağlık ekiplerinin kontrolleri ile toplanan gerçek hasta verilerini içeren önemli ve özgün bir çalışma niteliğindedir.

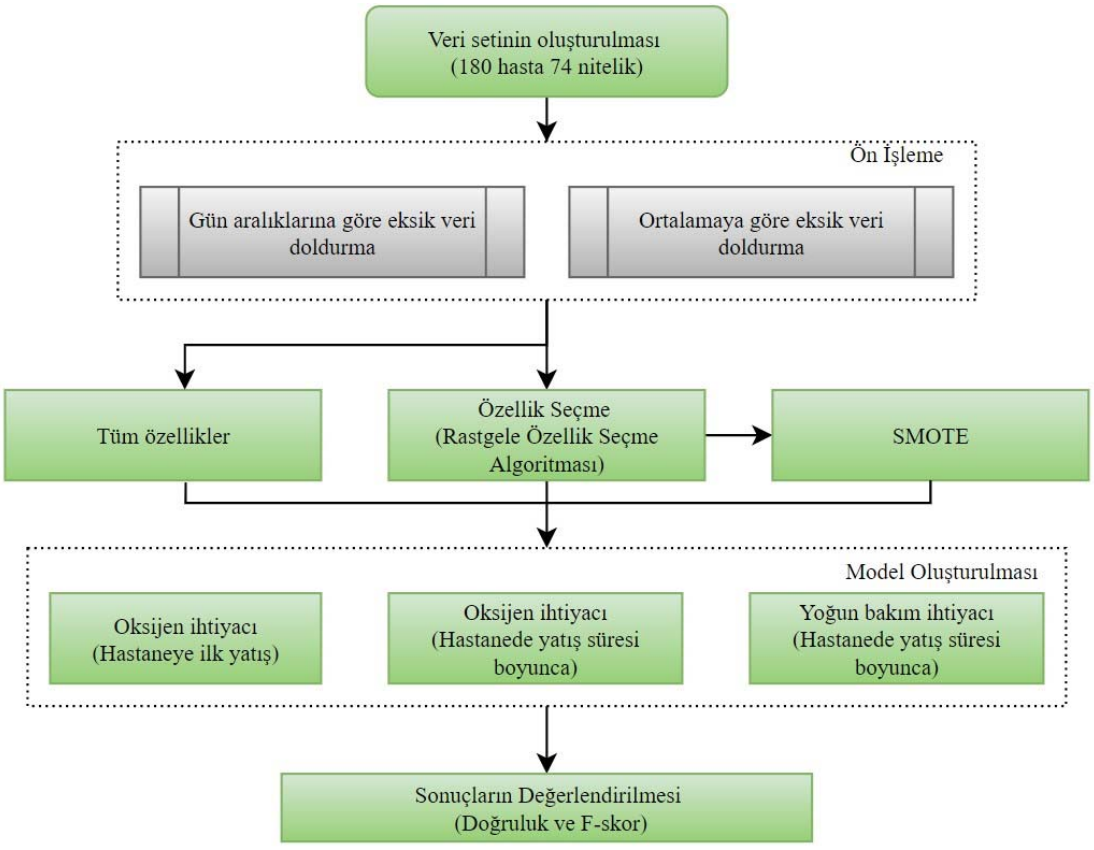
Aşağıdaki listelenen üç araştırma sorusu bu analizlere karşılık olarak çalışmanın başında oluşturulmuştur. Her bir araştırma sorusu SMOTE uygulanmamış dengesiz veri seti ve SMOTE uygulanmış veri seti üzerinde analiz edilmiştir.

- Hastanın hastaneye ilk yatış aşamasında oksijen tedavisine ihtiyacı olma olasılığı nedir?
- Hastanın hastanede yattığı süre boyunca oksijen tedavisine ihtiyacı olma olasılığı nedir?
- Hastanın hastanede yattığı süre boyunca yoğun bakım ihtiyacı olma olasılığı nedir?

Ayrıca tüm laboratuvar sonuçları yerine özellik seçme algoritması sonucunda elde edilen daha az sayıda özellik kullanılarak da yoğun bakım ve oksijen ihtiyacının yüksek doğruluk ile tespit edilebileceği gösterilmiştir ve analizler kapsamında yoğun bakım ve oksijen ihtiyacı ile ilgili en ilişkili özelliklerin bulunması hedeflenmektedir. Bu çalışmanın ikinci bölümünde kullanılan yöntemin işlem adımları ve kullanılan makine öğrenmesi yöntemleri, üçüncü bölümde bulgular ve son bölümde elde edilen sonuçlar tartışılmaktadır.

2. Deneysel Metot / Teorik Metot
(Experimental Method) / (Theoretical Method)

Çalışma veri setinin oluşturulması, önişleme, özellik seçme, model oluşturulması ve model sonuçlarının değerlendirilmesi olmak üzere



Şekil 1. Çalışmanın akış diyagramı (Flow chart of study)

toplamda 5 farklı temel aşamadan oluşmaktadır (Şekil 1). Çalışmanın ayrıntılı bir akış diyagramını temsil etmektedir.

2.1. Verisetinin Oluşturulması (Creating Dataset)

27 Nisan 2020 ve 1 Haziran 2020 tarihleri arasında İstanbul Marmara Üniversitesi Araştırma Hastanesi'nde COVID-19 için belirlenen bölümlere başvuran tüm hastaların 28 günlük takip süreçlerini içeren anonimleştirilmiş veriler bu çalışma kapsamında analiz edilmiştir. Tüm hastalara tıbbi öykü, mevcut şikayetler, yaşamsal belirtilerin ölçümü ve kapsamlı fizik muayene, kan sayımı için başlangıç testi, böbrek ve karaciğer fonksiyon testleri ve serum ferritin ve D-dimer dahil inflamatuar belirteçler dahil olmak üzere rutin klinik değerlendirmeler sağlık çalışanları tarafından yapılmıştır ve elde edilen bilgiler analiz edilmek üzere çalışmaya yerli bir veri seti çözümü olarak sunulmuştur. Çalışma kapsamında iki farklı veri seti oluşturularak bu iki veri setinin birleştirilmesi ile nihai veri seti elde edilmiştir ve analiz işlemleri yapılmıştır.

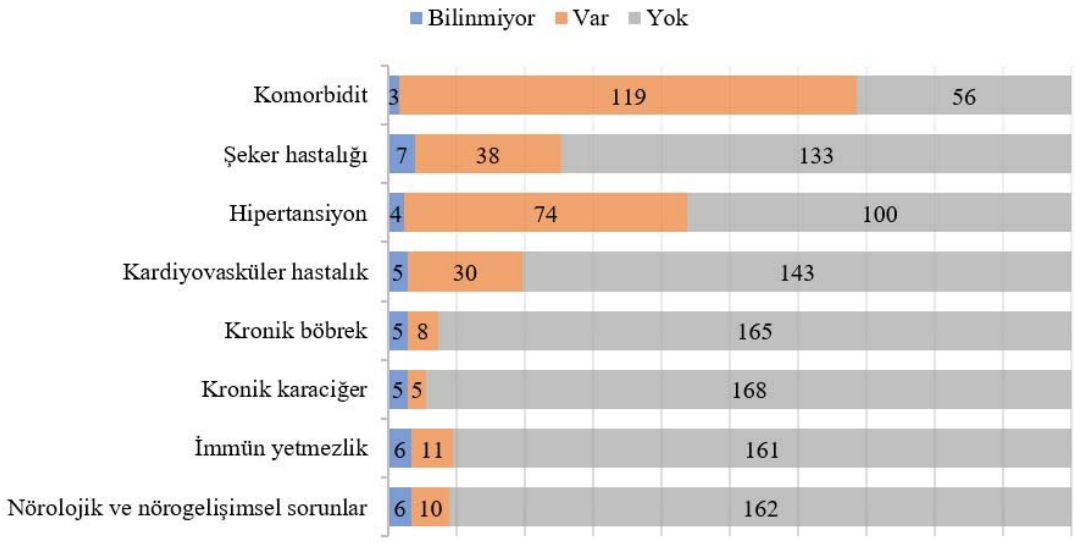
Bu çalışmada gerçekleştirilen tüm analiz işlemleri WHO ölçüt değeri hedef değer olarak kullanılarak gerçekleştirilmiştir [21]. WHO ölçüt değeri DSÖ (Dünya Sağlık Örgütü) liderliğinde kurulan Klinik Karakterizasyon ve Yönetim Çalışma Grubu tarafından geliştirilmiştir. Grup üyeleri epidemiyoloji, klinik araştırmalar, bulaşıcı hastalıklar ve viroloji bulaşıcı hastalıklar gibi alanlarda uzmanlığa sahiptir. WHO Klinik İlerleme Ölçeği, SARS-CoV-2 ile enfekte olmuş ve Covid-19'a yakalanmış hastalara fayda sağlamak amacıyla en uygun kaynak planlamasını sağlayabilmek ve bilgi alışverişini hızlandırabilmek için klinik araştırmalar ve kohort çalışmalar arasında veri havuzu oluşturma işlemini kolaylaştırmak üzere geliştirilmiştir [21]. Bu ölçüt değeri Covid-19 gibi ortaya çıkabilecek bir bulaşıcı hastalık salgınında kullanılmak üzere avantajlı bazı özellikler içermektedir. Klinik kayıtlardan elde edilerek oluşturulmuş bu değer 0 (enfekte olmayan) ile 10 (ölüm) arasında bir hastalık şiddeti ölçü bilgisi sağlar. Enfeksiyon olmama durumundan ölüm durumuna kadar uzanan bu geniş spektrum bu ölçeğin geniş bir aralıkta kullanılmasına imkân verir [21].

508 hasta kaydından oluşan laboratuvar veri seti ile 517 adet hasta kaydına ait tıbbi bilgileri içeren tıbbi bilgi veri seti T.C. kimlik numarası alanı kullanılarak birleştirilmiştir ve nihai veri seti elde edilmiştir. 508 hasta içerisinden 180 hastanın WHO giriş ve WHO en kötü değerleri bulunmaktadır bu yüzden kalan hastalar veri setinden çıkartılmıştır ve analize dahil edilmemiştir. Veri setindeki hastaların ortanca yaş değeri (medyan) 55 ve yaş ortalaması yaklaşık 56,01'dir. Hastaların 85 (%47,2) tanesi kadın iken geri kalan 95 (%52,8) hastanın cinsiyeti erkektir (Şekil 2). Veri setinde bulunan 180 adet hastanın tıbbi öyküsünü göstermektedir. Eştanı (komorbidite), belirli zaman diliminde birden fazla hastalığın birlikte bulunmasıdır. Kardiyovasküler hastalıklar kapsamında KAH, SVO, periferik VH ve immün yetmezlik kapsamında HIV, kanser, otoimmün gibi hastalıklar bulunmaktadır. Olguların %67,2'sinde en sık hastaneye yatış nedeni (%45,0), ardından ileri yaş (%37,7), dispne/hipoksi (%35,5) ve şiddetli pnömoninin radyolojik kanıtı (%34,4) olan komorbiditeler olarak belirlenmiştir.

Çalışma kapsamında tüm analiz işlemleri WHO ölçüt değeri hedef değer olarak kullanılarak gerçekleştirilmiştir. Klinik Karakterizasyon ve Yönetim Çalışma Grubu tarafından geliştirilen WHO Klinik İlerleme Ölçeği COVID-19 hastalarına en uygun kaynak planlaması sağlanması, hızlı bilgi aktarımı ve klinik araştırma veri havuzu oluşturma işlemini kolaylaştırmak üzere geliştirilmiştir. Klinik kayıtlara bağlı olarak elde edilen bu değer 0 (enfekte olmayan) ile 10 (ölüm) arasında bir hastalık şiddeti ölçü bilgisi sağlamakta ve bu 10 farklı WHO değeri kullanılarak hastalar beş farklı sınıfta değerlendirilmektedir [21].

0 değeri hasta için herhangi bir enfekte olmama durumunu ve 10 değeri ise ölüm durumunu temsil etmektedir. Kalan WHO değerleri ise ayakta hafif hastalık (Ambulatory mild disease), hastanede yatış: orta derecede hastalık (Hospitalised: moderate disease) ve hastanede yatış: ciddi derecede hastalık (Hospitalised: severe disease) olmak üzere üç farklı başlık altında değerlendirilebilir [21].

Analizlerin amacına göre WHO giriş veya WHO en kötü değerleri veri setinde hedef değer olarak kullanılmıştır (Şekil 3). 180 hastanın WHO giriş-en kötü değerlerinin dağılımını göstermektedir. WHO Giriş hastanın hastaneye yatış aşamasındaki hastalık şiddetini temsil eder ve "Hastanın hastaneye ilk yatış aşamasında oksijen tedavisi ihtiyacı olma olasılığı nedir?" analizinde hedef değer olarak kullanılmıştır. WHO en kötü ise hastanın hastanede yatış süresi boyunca ulaştığı en yüksek hastalık şiddetini temsil eder ve "Hastanın hastanede yattığı süre boyunca oksijen tedavisi ihtiyacı olma olasılığı nedir?" ve "Hastanın hastanede yattığı süre boyunca yoğun bakım



Şekil 2. Hastaların tıbbi öyküsü (Medical history of the patients)

ihtiyacı olma olasılığı nedir?" analizlerinde hedef değer olarak kullanılmıştır.

2.2. Ön İşleme (Preprocessing)

Veri seti gerçek dünya verilerinden oluşmaktadır ve benzersiz bir veri seti olduğundan eksik veriler içermektedir. Veri madenciliği işlemlerinde uygulanan eforun yaklaşık %80'i veri kalitesinin artırılmasına harcanmaktadır [22]. Veri ön işleme aşaması veri temizleme, veri entegrasyonu, veri seçimi, veri dönüştürme vb. gibi etkileşimli adımlardan oluşur [23]. Gerçekleştirilen çalışma kapsamında veri seti içerisinde yer alan eksik veriler gün aralıklarına göre ve ortalama değer atama yöntemleri kullanılarak tamamlanmıştır. İlk olarak özellikler için belirlenen gün aralıklarına göre eksik veri doldurma işlemi yapıldıktan sonra eksik veriler içeren özellikler ortalama değer yöntemi kullanılarak tamamlanmıştır. Enfeksiyon hastalıkları uzman doktorları ile belirlenen karar çerçevesinde gerçek laboratuvar sonuçları ile en iyi örtüşme sağlayacak şekilde bir değer doldurma yapılması uygun görüldüğü için bu yöntem kullanılmıştır

2.2.1. Gün aralıklarına göre eksik veri doldurma
(Filling in missing data according to day intervals)

Alanında uzman sağlık çalışanlarından alınan fikirler doğrultusunda uygulanan gün aralıklarına göre eksik veri doldurma tekniği veri setine ilk olarak uygulanan veri doldurma tekniğidir. Belirlenen özellikler için gün aralık değerleri, alanında beş yıldan fazla süredir uzman doktor olarak çalışan enfeksiyon hastalıkları uzman doktorları tarafından belirlenerek bu çalışmada kullanılmıştır. Bu eksik veri doldurma tekniği kapsamında sağlık çalışanları yardımı ile 185 laboratuvar sonucundan 75 tanesi için gün aralıkları tanımlanmıştır. Kalan 110 özellik ise alt laboratuvar sonuçları olduğundan veya bir dönem bakıldığından dolayı çok eksik veri içerdiği için uzman görüşü de dikkate alınarak veri setinden çıkartılarak analizlere dahil edilmemiştir. Analizlerde sadece hastanın ilk laboratuvar sonucu kullanıldığından dolayı hastanın diğer laboratuvar sonuçları sağlık çalışanları tarafından belirlenen gün aralıkları ile birlikte kullanılarak hastanın ilk laboratuvar sonucundaki eksik veri oranının düşürülmesi hedeflenmiştir (Şekil 4). Örnek bir laboratuvar değeri için oluşturulan gün aralıklarına göre eksik veri doldurma örneğini temsil etmektedir.

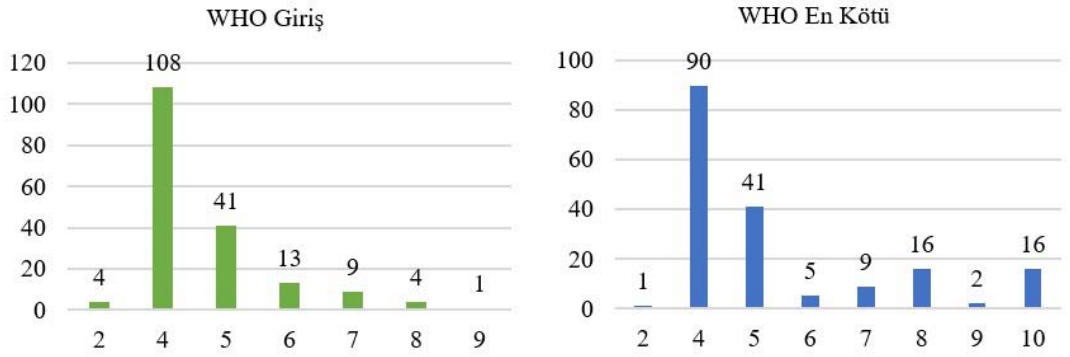
2.2.2. Ortalama değer atama yöntemi kullanılarak eksik veri doldurma
(Fill the missing data using the mean value)

Ortalama değer atama yöntemi, her eksik değeri özniteliğin ait olduğu sınıfın ortalamasıyla değiştirir. Yalnızca sayısal nitelikler için ve örneklerin önceden sınıflandırıldığı, verinin sınıflara ayrıldığı veri setlerinde kullanılabilir [24]. Eş. 1'de görüldüğü gibi nitelikteki eksik veri, verinin sahip olduğu sınıf ile aynı değere ait olan ve boş olmayan değerlerin toplamının veri sayısına oranı kullanılarak doldurulur.

$$\text{Ortalama değer} = \frac{\text{Eksik veri ile aynı sınıfa ait değerlerin toplamı}}{\text{Veri sayısı}} \quad (1)$$

2.3. Özellik Seçimi (Feature Selection)

Özellik altkümesi seçimi, veri madenciliği sürecinde çok önemli bir rol oynamaktadır [25-27]. Bu çalışmada RSFS (Random Subset Feature Selection) algoritması kullanılarak veri setindeki öznitelik



Şekil 3. WHO giriş-en kötü değerlerinin dağılımı (Distribution of WHO initial-worst values)

Albumin gün aralığı : 5

Ad Soyad	Sonuç Alınma Tarihi	Albumin
Hasta - 1	2020-04-02 2:20	42
Hasta - 1	2020-04-02 2:31	
Hasta - 1	2020-04-02 2:37	42
Hasta - 1	2020-04-02 10:34	45

Şekil 4. Gün aralıklarına göre eksik veri doldurma örneği (Example of filling in missing data according to day intervals)

sayısı azaltılmıştır [28]. Bu sayede daha az öznitelik kullanılarak daha yüksek doğruluk değerleri elde edilmiştir. RSFS algoritması üç ana başlıktan oluşmaktadır. Bunlar; ön işleme, rasgele altküme öznitelik seçimi ve sınıflandırma. Ön işleme aşamasından önce, tüm özellikler 0 ortalama ve birim standart sapmaya sahip olacak şekilde z-skor dönüşümü ile normalize edilir. Bu çalışmada özellik altkümesi seçimi için kullanılan RFSS yöntemi için niteliklerin tümü z-skor dönüşümü yapılarak uygulamalar gerçekleştirilmiştir.

RSFS; Özellik seçme işlemi yinelemeli olarak yapılır. Rastgele seçilen özellik alt kümeleri, KNN (k-nearest neighbors) sınıflandırıcısı tarafından sınıflandırılır. Her adımda, rastgele seçilen özellikler ilgililik değerlerine göre derecelendirilir. Her bir alt kümede, toplam özellik sayısının karekökü alınarak elde edilen değer ile rastgele seçilen özellikler vardır.

Sınıflandırma: Eğitim veri seti, rastgele oluşturulmuş alt kümedeki özellikler kullanılarak KNN sınıflandırıcısı ile sınıflandırılır. Rastgele oluşturulmuş her bir alt kümenin ilgililik değeri, P.C (performans kriteri) ve E.C beklenen kriteri arasındaki fark olarak hesaplanır. E.C, tüm yinelemelerdeki geri çağırma değerlerinin ortalamasıdır (Eş. 2) kullanılarak hesaplanır.

$$E(c) = \sum \frac{\text{Doğru sınıflandırılanlar}}{(\text{Doğru sınıflandırılanlar})+(\text{Yanlış sınıflandırılanlar})} x \frac{100}{n} \quad (2)$$

P.C, mevcut yinelemeler için geri çağırma değerinin ortalamasıdır. Eş. 3 kullanılarak hesaplanır.

$$\text{İlişki düzeyi (r)} = (\text{P. C(performans kriteri)}) - (\text{E. C(beklenen kriter)}) \quad (3)$$

2.4. Yapay Veri Çoğaltma (Synthetic Data Augmentation)

SMOTE dengesiz veri dağılımına sahip olan veri kümelerinde azınlık olan sınıfın yapay bir şekilde arttırılmasını sağlayan yapay veri çoğaltma yöntemidir [29].
Algoritma çalışma adımları aşağıdaki gibidir [30]:

- Azınlık sınıftaki bir örneğin en yakın k adet komşusu Öklid mesafesi kullanılarak bulunur.
- K adet komşudan rastgele biri seçilir.
- A ve B noktaları birleştirilerek özellik uzayında bir çizgi oluşturularak sentetik veri elde edilir. Oluşturulan sentetik veriler A ve B örneğinin dışbükey kombinasyonunu ifade eder.

2.5. Sınıflandırma (Classification)

Çalışmada KNN, Bagging, Rastgele Orman ve Karar Ağacı olmak üzere 4 farklı makine öğrenmesi sınıflandırma yöntemi kullanılmıştır. Sınıflandırma işlemi Weka uygulaması üzerinde gerçekleştirilmiştir. Weka uygulaması makine öğrenmesi algoritmaları ve veri önişleme işlemleri için sıklıkla kullanılan Waikato üniversitesi tarafından açık kaynak kodlu olarak geliştirilmiş bir veri madenciliği programıdır [31].

Tüm sınıflandırıcı doğrulukları sınıflandırma doğruluğu değeri kullanılarak değerlendirilmiştir ve tüm doğruluk değerleri K katlamalı çapraz doğrulama tekniği kullanılarak elde edilmiştir. K katlamalı çapraz doğrulama veri madenciliğinde oluşturulan modelin başarısını sınamak için kullanılan yöntemlerden birisidir. Literatürde sıklıkla kullanılan k değerinin 10 olması da temel alınarak bu çalışmada yapılan tüm analizlerde k değeri 10 olarak alınmıştır [32]. Doğruluk değerinin yanında F1 skor değeri de çalışma kapsamında kullanılan metrikler arasındadır. F1 skorunun dengesiz bir dağılıma sahip olan veri setlerinde kullanılması daha uygundur. Çalışmada kullanılan veri seti de dengesiz bir dağılıma sahip olduğundan dolayı farklı sınıfların doğruluklarını da değerlendirmek için bu iki metrik incelenmiştir.

2.5.1. K-en yakın komşu (KNN)

KNN sınıflandırma algoritması basit ve temel sınıflandırma yöntemlerindendir. Parametrik olmayan bir yöntem olan KNN algoritması regresyon ve sınıflandırma için kullanılır. Algoritma çıktısı olarak bir sınıf etiketi üretilir. Sınıflandırma işlemi çoğunluk oylaması kullanılarak hesaplanır. Tahmin edilen nesne en yakın k komşusu arasında en fazla olan sınıfa atanır. Uzaklık değeri ve kullanıcı tanımlı pozitif k değeri olmak üzere iki değer kullanılarak tahmin işlemi gerçekleştirilir. Uzaklık değeri farklı uzaklık ölçme formülleri kullanılarak hesaplanabilir. Bu çalışmada Öklid Uzaklığı (Euclidean Distance) formülü mesafeyi ölçmek için kullanılmıştır. Öklid uzaklığı noktalar arasındaki doğrusal uzaklığı hesaplanan için kullanılan bir uzaklık hesaplamasıdır ve formülü Eş. 4'de verilmiştir [32].

$$d(a, b) = \sqrt{\sum_{i=1}^{n} (a1 - b1)^2} \quad (4)$$

2.5.2. C4.5 Karar ağacı (Decision tree)

İlk olarak 1960 yılında tanıtılan karar ağacı algoritması yaygın olarak kullanılan bir makine öğrenmesi yöntemidir. Yöntem verisetindeki bilgileri kullanarak bir kök düğüm, iç düğümler ve yaprak düğümlerden oluşan bir ağaç modeli oluşturmaktadır. Algoritma parametrik değildir ve karmaşık verisetlerinde parametre kullanımı olmadan verimli sonuçlar elde edebilir [33].

Çalışmada C4.5 karar ağacı algoritması kullanılmıştır. Bu algoritmanın Weka yazılımındaki karşılığı J48'dir. J48 algoritması azaltılmış hata budaması ve açgözlü yöntemi kullanarak karar ağacını oluşturur. Temelinde bilgi entropisi kavramını kullanır ve entropi farkını inceler, bu entropi farkı bilgi kazancı olarak adlandırılır. Temelinde işlem adımları aşağıdaki gibidir [31].

- Örnekler aynı sınıfa ait ise bu düğüm sınıf etiketi belirlenmiş tek yaprağı ifade eder. Aksi durumda özniteliğe göre bölerek kara ağacı oluşturulur ve tüm örnekler aynı sınıfa ait ise bölme işlemi sonlandırılır.
- Her bir nitelik için bilgi kazancı değeri hesaplanır.
- En yüksek bilgi kazancına sahip nitelik kullanılarak ağaç kırılımları oluşturulur.

2.5.3. Rastgele orman (Random forest)

Rastgele Orman algoritması toplulukla öğrenme yöntemlerinden torbalama tekniğine örnek olarak verilebilir. Torbalama algoritması rastgele olarak oluşturulan veri setleri üzerinde paralel olarak sınıflandırma analizi yapan bir makine öğrenmesi modelidir. Temelinde karar ağaçları yer alan bir makine öğrenmesi modeli olan Rastgele Orman algoritmasında veri üzerinde aşırı öğrenme probleminin üstesinden gelinmesini sağlayan bir yapı mevcuttur. Temel olarak karar ağaçlarının aksine birden fazla karar ağacı yapısı oluşturularak daha yüksek bir sınıflandırma değerine ulaşılması hedeflenmiştir.

Rastgele Orman algoritması veriseti N adet kayıta ve M adet özelliğe sahip olduğunda her bir ağaç yapısı için N adet kayıt seçilerek ve m<M olacak şekilde m adet özellik seçilerek yapılır. Her bir karar ağacı yapısı m özellikli N adet kayda sahip alt verisetini kullanarak bir karar ağacı modeli oluşturur. Böylece her bir karar ağacı farklı alt eğitim setlerini kullanarak model oluşturur. Test aşamasında ise verilen örnek her bir ağaç tarafından tek tek değerlendirilir ve nihai sınıflandırma çoğunluk oylaması kullanılarak belirlenir. Böylece

birkaç zayıf karar ağacı sınıflandırıcısı birleştirilerek güçlü bir sınıflandırıcı elde edilmiş olur [34, 35].

2.5.4. Torbalama (Bagging)

Bootstrap toplama (Bootstrap aggregating) olarak da bilinen Torbalama (Bagging) sınıflandırıcısı ilk olarak Breiman tarafından 1996 yılında tanıtılmıştır. Torbalama sınıflandırıcısı nesnelerin örneklerini alır her bir örnek üzerinde bir sınıflandırıcı geliştirir. Çoğunluk oylama tekniği kullanılarak sınıflandırıcı oyları birleştirilir [36].

Weka kütüphanesinde DecisionStump, J48, LMT, RepTree, RandomTree olmak üzere farklı torbalama yöntemleri mevcuttur. Rastgele Ağaç (RandomTree) yöntemi, rastgele seçilen K özniteliklerini dikkate alarak budama yapmadan bir ağaç oluşturur. Ek olarak, ağaçlardaki her parametre, örneğin yaprak sayısı veya çap gibi rastgele bir değişkendir. RandomTree ayrıca sınıf olasılıklarının tahminine izin verme seçeneğine veya bir gerileme setine dayalı olarak regresyon durumunda hedef ortalamaya izin verme seçeneğine sahiptir [37].

3. Bulgular (Results)

Hastaneye ilk kabul aşamasında oksijen ihtiyacının tahmin edilmesi, hastanede yatış süresi boyunca oksijen ihtiyacı ve yoğun bakım ihtiyacının tahmin edilmesi olmak üzere 3 farklı analiz, model oluşturma işlemi gerçekleştirilmiştir. Hastaneye ilk kabul aşamasında tahmin işlemi yapan analizlerde hedef değer olarak WHO giriş kullanılırken hastanede yatış süresi boyunca tahmin işlemi yapan analizlerde hedef değer olarak WHO en kötü kullanılmıştır. Her bir uygulama için kullanılan 4 farklı sınıflandırma algoritması doğruluk, F1 skoru metrikleri açısından değerlendirilmiştir. Oluşturulan verisetinin 10-çapraz katlama yöntemi ile kullanılarak yapılan analizlere ait en iyi performansa sahip olan algoritma seçilmiştir. Çalışmadaki veriseti sınıflama yöntemlerinin uygulanması sırasında 10 çapraz doğrulama yöntemi ile kullanılmıştır. Ayrıca yapay bir veri çoğaltma tekniği olan SMOTE uygulandıktan sonra oluşturulan modellerinde sonuçları karşılaştırılarak SMOTE algoritmasının sınıflandırma performansına katkısı değerlendirilmiştir. Her analiz işleminde önce ortalamaya göre eksik veri doldurma tekniği kullanılarak eksik veriler tamamlanmıştır.

Tablo 1. Her bir analizde kullanılan toplam hasta sayısını, belirlenen hedef değeri ve sınıf dağılımını göstermektedir.

3.1. Hastaneye İlk Kabul Aşamasında Oksijen İhtiyacının Tahmin Edilmesi
(Estimation of Oxygen Demand During the Initial Admission to Hospital)

Bu analizde Tablo 1'de "Hastaneye ilk yatışta oksijen ihtiyacı tahmini" olarak etiketlenen veri seti kullanılmıştır. Veri seti toplamda 180 adet hastadan ve 72 nitelikten oluşmaktadır. Hastalardan 95 tanesi erkek iken kalan 85 hastanın cinsiyeti kadındır ve yaş ortalaması yaklaşık 56,01'dir.

Hastaneye ilk yatış aşamasındaki oksijen ihtiyacının tahmin edilmesi hedeflendiği için WHO giriş hedef olarak kullanılmıştır ve ortalama eksik veri doldurma tekniği kullanılarak eksik veriler model oluşturma işleminde başlamadan önce doldurulmuştur. Ayrıca her analiz için oluşturulan veri setine Rastgele Özellik Seçme algoritması uygulanarak analiz ile en ilişkili özellikler belirlenmiştir. Tablo 2 analiz için kullanılan 72 adet niteliğin listesini ve tüm özellikler arasından Rastgele Özellik Seçme algoritması kullanılarak elde edilen 16 niteliği temsil etmektedir.

WHO giriş için sınıf dağılımını belirleyecek olan hedef değer 4 olarak belirlenmiştir. 180 adet hastadan WHO giriş değeri dört ve altında olan hastalar (112) "0" olarak etiketlenirken kalan hastalar (68) ise "1" olarak etiketlenmiştir. 4 farklı makine öğrenme tekniği kullanılarak Weka üzerinde oluşturulan modellerin doğruluk ve F1-skor değerleri karşılaştırılmıştır. Tablo 3. tüm özellikler kullanılarak ve sadece Rastgele Özellik Seçme algoritması sonucunda en ilişkili olarak belirlenen özellikler kullanılarak elde edilen modellerin doğruluk karşılaştırmasını temsil etmektedir. Sonuçlar %91,67 genel doğruluk performansı ile 16 özelik kullanılarak hastaneye yatış

Tablo 1. Analize göre kullanılan veri seti bilgileri (Data set information used according to the analysis)

Analiz Numarası	Veri seti	Hedef Değer	Veri seti dağılımı (Sınıf 0 – Sınıf 1)	Toplam hasta sayısı
1	Hastaneye ilk yatışta oksijen ihtiyacı tahmini	WHO Giriş	112 - 68	180
2	Hastanede yatış süresi boyunca oksijen ihtiyacı tahmini	WHO en kötü	91 - 21	112
3	Hastanede yatış süresi boyunca yoğun bakım ihtiyacı tahmini	WHO en kötü	137 - 29	166

Tablo 2. Analiz – 1 Tüm özellikler ve seçilen özellikler (Analysis – 1 All features and selected features)

Özellikler				Seçilen Özellikler
Yaş	EOS#	Direkt Bilirubin	MCHC	CRP (Nefelometrik)
Cinsiyet	EOS%	Karboksihemoglobin	MCV	ALP
NEU%	Laktat	Kreatinin	PCT	Ferritin
Sodyum	PLT	Kütle CK-MB	MCH	Total Protein
ALT	Fosfor	Total Bilirubin	MON#	D-dimer
aPTT	WDOP	Baz Açığı	MON%	Troponin T-hs
AST	WBC	Prokalsitonin	MPV	Ürik asit
BAS#	HCO3-	Ozmolarite	NRBC#	Methemoglobin
BAS%	HCT	Magnezyum		GGT
LDH	HGB	plt/lymph#		Glukoz
BUN	PDW	neu#/lymph#		Fibrinojen
Klor	pO2	Kalsiyum		lymph#/crp
PT%	Üre	LYMPH#		Albumin
RBC	SO_2	LYMPH%		MDW
INR	NEU#	Potasyum		pCO_2
RDW	PT	NRBC		pH

aşamasındaki oksijen ihtiyacının yüksek bir sınıflandırma performansı ile tespit edilebildiğini göstermiştir.
Tablo 4. 72 özellik ve 16 özellik kullanılarak oluşturulan en yüksek sınıflandırma doğruluk değerine sahip modellerin F-Skor değerlerini göstermektedir. 16 özellikli Rastgele Orman modeli oksijen tedavisine ihtiyaç duyan hastaları 0,933 F1-Skor ile sınıflandırırken ihtiyaç duymayan hastaları 0,889 F1-Skoru ile sınıflandırmıştır ve sınıf bazında da iyi bir sınıflandırma performansı elde edilmiştir.

Tablo 5. 16 özellik kullanılarak ve SMOTE uygulandıktan sonra 16 özellik kullanılarak oluşturulan model sonuçlarının doğruluk ve F1-skor karşılaştırılmasını temsil etmektedir. %96,64 değeri ile SMOTE tekniği uygulandıktan sonra genel doğruluk değeri açısından yaklaşık %5'lik bir artış görülmüştür. SMOTE kullanılmadan önce 0,89 F1-Skoruyla tahmin edilen oksijen tedavisine ihtiyaç duyan hastalar SMOTE uygulandıktan sonra 0,95 F1-Skor değerine yükselmiştir.

Şekil 5. sırayla 16 özellikli SMOTE uygulanmayan ve uygulanan modellerin karmaşıklık matrislerini temsil etmektedir. Karmaşıklık matrislerine göre SMOTE uygulanmadan önce 1 etiketli oksijen tedavisine ihtiyaç duyan sekiz hasta yanlış sınıflandırılırken SMOTE uygulandıktan sonra bu değer dörde düşmüştür.

3.2. Hastanede Yatış Süresi Boyunca Oksijen İhtiyacının Tahmin Edilmesi
(Estimation Of Oxygen Demand During Hospitalization)

Bu analizde Tablo 1'de "Hastanede yatış süresi boyunca oksijen ihtiyacı tahmini" olarak etiketlenen veri seti kullanılmıştır. Veri seti toplamda 112 adet hastadan ve 73 nitelikten oluşmaktadır. Hastalardan 54 tanesi erkek iken kalan 58 hastanın cinsiyeti kadındır ve yaş ortalaması yaklaşık 51,72'dir. Analize başlamadan önce hastaneye yatış aşamasında çoktan oksijen tedavisine ihtiyaç duyan hastalar (68) veri setinden çıkartılarak kalan 112 hasta analize dâhil

Tablo 3. Analiz – 1 Tüm özellikler ve seçilen özellikler ile oluşturulan modellerin doğruluk ve parametre değerleri
(Analysis – 1 performance and parameters of models created with all features and selected features)

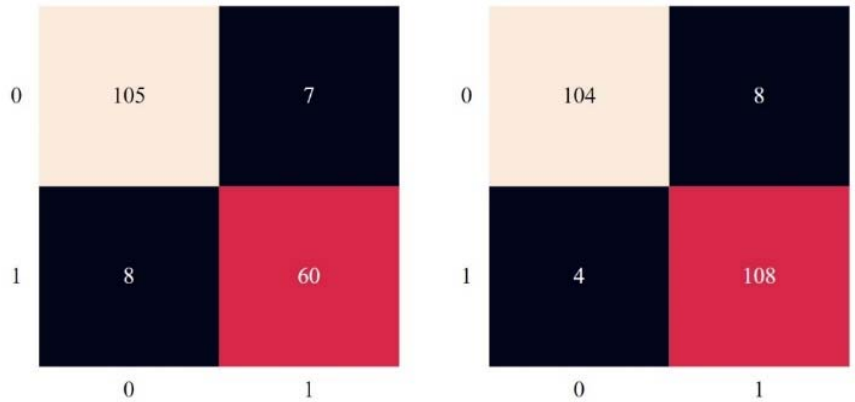
Özellik Sayısı	Algoritma	Doğruluk	Parametre
72	KNN	%73,89	K: 7
	Bagging	%87,22	Sınıflandırıcı: RandomTree
	Rastgele Orman	*%89,44*	*Ağaç sayısı: 100*
	Karar Ağacı	%85	(Minimum obje sayısı: 5, Güven faktörü: 0,1)
16	KNN	%82,22	K: 5
	Bagging	%88,33	Sınıflandırıcı: RandomTree
	Rastgele Orman	*%91,67*	*Ağaç sayısı: 40*
	Karar Ağacı	%87,22	(Minimum obje sayısı: 4, Güven faktörü: 0,25)

Tablo 4. Analiz – 1 Özellik sayısına göre en iyi performansa sahip modellerin F-Skor değeleri
(Analysis – 1 performance of models created with all features and selected features)

Model Adı	Doğruluk	Oksijen tedavisine ihtiyaç	F1-Skor
72 özellik Karar Ağacı	%89,44	VAR	0,853
		YOK	0,918
16 özellik Rastgele Orman	*%91,67*	VAR	*0,889*
		YOK	*0,933*

Tablo 5. Analiz – 1 SMOTE uygulanan ve uygulanmayan modellerin performansları
(Analysis – 1 performance of models with and without SMOTE)

Model Adı	Doğruluk	Oksijen tedavisine ihtiyaç	F1-Skor
16 özellik Rastgele Orman	%91,67	VAR	0,89
		YOK	0,93
SMOTE - 16 özellik Rastgele Orman	*%96,64*	VAR	*0,95*
		YOK	*0,95*



Şekil 5. Analiz – 1 SMOTE uygulanmayan ve uygulanan modellerin karmaşıklık matrisi
(Analysis – 1 Confusion matrix of non-SMOTE and applied models)

edilmiştir. Hastanede yatış süresi boyunca oksijen ihtiyacının tahmin edilmesi hedeflendiği için WHO en kötü hedef olarak kullanılmıştır ve ortalama eksik veri doldurma tekniği kullanılarak eksik veriler model oluşturma işleminde başlamadan önce doldurulmuştur. Ayrıca her analiz için oluşturulan veri setine Rastgele Özellik Seçme algoritması uygulanarak analiz ile en ilişkili özellikler belirlenmiştir. Tablo 6. tüm özellikler arasından Rastgele Özellik Seçme algoritması kullanılarak elde edilen 18 niteliği temsil etmektedir.

WHO en kötü için sınıf dağılımını belirleyecek olan hedef değer 4 olarak belirlenmiştir. 112 adet hastadan WHO en kötü değeri dört ve altında olan hastalar (91) "0" olarak etiketlenirken kalan hastalar (21) ise "1" olarak etiketlenmiştir. 4 farklı makine öğrenme tekniği kullanılarak Weka üzerinde oluşturulan modellerin doğruluk ve F1-skor değerleri karşılaştırılmıştır. Tablo 7 tüm özellikler kullanılarak ve sadece Rastgele Özellik Seçme algoritması sonucunda en ilişkili olarak belirlenen özellikler kullanılarak elde edilen modellerin doğruluk karşılaştırmasını temsil etmektedir. Sonuçlar %91,96 genel doğruluk performansı ile 18 özelik kullanılarak hastaneye yatış aşamasındaki oksijen ihtiyacının yüksek bir sınıflandırma performansı ile tespit edilebildiğini göstermiştir.

Tablo 8 73 özellik ve 18 özellik kullanılarak oluşturulan en yüksek sınıflandırma doğruluk değerine sahip modellerin F-Skor değerlerini göstermektedir. 18 özellik kullanılarak oluşturulan Rastgele Orman modeli oksijen tedavisine ihtiyaç duyan hastaları 0,952 F1-Skor ile sınıflandırırken veri sayısının az olmasından dolayı oksijen tedavisine ihtiyaç duymayan hastaları 0,743 F1-Skoru ile ortalama bir sınıflandırma performansıyla tahmin etmiştir. Tablo 9 18 özellik kullanılarak ve SMOTE uygulandıktan sonra 18 özellik kullanılarak oluşturulan model sonuçlarının doğruluk ve F1-skor karşılaştırılmasını temsil etmektedir. %97,8 değeri ile SMOTE tekniği uygulandıktan sonra genel doğruluk değeri açısından yaklaşık %6'lık bir artış görülmüştür. SMOTE kullanılmadan önce 0,74 F1-Skoruyla tahmin edilen oksijen tedavisine ihtiyaç duyan hastalar SMOTE uygulandıktan sonra 0,98 F1-Skor değerine yükselmiştir.

Şekil 6 sırayla 18 özellikli SMOTE uygulanmayan ve uygulanan modellerin karmaşıklık matrislerini temsil etmektedir. Karmaşıklık matrislerine göre SMOTE uygulanmadan önce "1" etiketli oksijen tedavisine ihtiyaç duyan sekiz hasta yanlış sınıflandırılırken SMOTE uygulandıktan sonra bu değer ikiye düşmüştür. Yanlış sınıflandırılan "1" etiketli hastalardan biri yapay bir veri iken ve diğeri gerçek bir hasta kaydına aittir.

3.3. Hastanede Yatış Süresi Boyunca Yoğun Bakım İhtiyacının Tahmin Edilmesi
(Estimating The Need For Intensive Care During Hospitalization)

Bu analizde Tablo 1'de "Hastanede yatış süresi boyunca yoğun bakım ihtiyacı tahmini" olarak etiketlenen veri seti kullanılmıştır. Veri seti toplamda 166 adet hastadan ve 73 nitelikten oluşmaktadır. Hastalardan 84 tanesi erkek iken kalan 82 hastanın cinsiyeti kadındır ve yaş ortalaması yaklaşık 55,03'dir. Analize başlamadan önce hastaneye yatış aşamasında çoktan yoğun bakım tedavisine ihtiyaç duyan hastalar (14) veri setinden çıkartılarak kalan 166 hasta analize

Tablo 6. Analiz – 2 seçilen özellikler (Analysis – 2 selected features)

Seçilen Özellikler						
Albumin	CRP (Nefelometrik)	Direkt Bilirubin	GGT	INR	Kalsiyum	LDH
LYMPH%	MDW	MON%	NEU	NEU#	PO2	Total Protein
Troponin T-hs	WBC	WDOP				

Tablo 7. Analiz – 2 Tüm özellikler ve seçilen özellikler ile oluşturulan modellerin doğruluk ve model parametre değerleri (Analysis – 2 performance and parameters of models created with all features and selected features)

Özellik Sayısı	Algoritma	Doğruluk	Parametre
73	KNN	%82,14	K: 4
	Bagging	%87,5	Sınıflandırıcı: DecisionStump
	Rastgele Orman	%87,5	Ağaç sayısı: 80
	Karar Ağacı	*%88,39*	(Minimum obje sayısı: 4, Güven faktörü: 0,25)
18	KNN	%84,82	K: 6
	Bagging	%87,5	Sınıflandırıcı: DecisionStump
	Rastgele Orman	*%91,96*	Ağaç sayısı: 60
	Karar Ağacı	%87,5	(Minimum obje sayısı: 3, Güven faktörü: 0,25)

Tablo 8. Analiz – 2 Tüm özellikler ve seçilen özellikler ile oluşturulan modellerin performansları (Analysis – 2 performance of models created with all features and selected features)

Model Adı	Doğruluk	Oksijen tedavisine ihtiyaç	F1-Skor
73 özellik C4.5 Karar Ağacı	%88,39	VAR	0,698
		YOK	0,928
18 özellik Rastgele Orman	***%91,96***	VAR	*0,743*
		YOK	*0,952*

Tablo 9. Analiz – 2 SMOTE uygulanan ve uygulanmayan modellerin performansları (Analysis – 2 performance of models with and without SMOTE)

Model Adı	Doğruluk	Oksijen tedavisine ihtiyaç	F1-Skor
18 özellik Rastgele Orman	%91,96	VAR	0,74
		YOK	0,95
SMOTE - 18 özellik Rastgele Orman	*%97,8*	VAR	0,98
		YOK	0,98

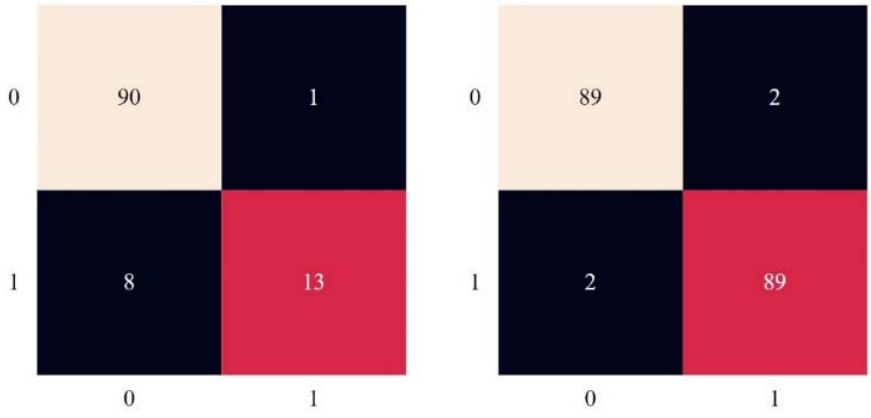
dahil edilmiştir. Hastanede yatış süresi boyunca yoğun bakım ihtiyacının tahmin edilmesi hedeflendiği için WHO en kötü hedef olarak kullanılmıştır ve ortalama eksik veri doldurma tekniği kullanılarak eksik veriler model oluşturma işleminde başlamadan önce doldurulmuştur.

Ayrıca her analiz için oluşturulan veri setine Rastgele Özellik Seçme algoritması uygulanarak analiz ile en ilişkili özellikler belirlenmiştir. Tablo 10 tüm özellikler arasından Rastgele Özellik Seçme algoritması kullanılarak elde edilen 12 niteliği temsil etmektedir.

WHO en kötü için sınıf dağılımını belirleyecek olan hedef değer 6 olarak belirlenmiştir. 166 adet hastadan WHO en kötü değeri altı ve altında olan hastalar (137) "0" olarak etiketlenirken kalan hastalar (29) ise "1" olarak etiketlenmiştir. 4 farklı makine öğrenme tekniği kullanılarak Weka üzerinde oluşturulan modellerin doğruluk ve F1-skor değerleri karşılaştırılmıştır. Tablo 11. tüm özellikler kullanılarak ve sadece Rastgele Özellik Seçme algoritması sonucunda en ilişkili olarak belirlenen özellikler kullanılarak elde edilen modellerin doğruluk karşılaştırmasını temsil etmektedir. Sonuçlar %92,17 genel doğruluk performansı ile 12 özelik kullanılarak hastanede yatış süresi boyunca yoğun bakım ihtiyacının yüksek bir sınıflandırma performansı ile tespit edilebildiğini göstermiştir.

Tablo 12 73 özellik ve 12 özellik kullanılarak oluşturulan en yüksek sınıflandırma doğruluk değerine sahip modellerin F-Skor değerlerini göstermektedir. 73 özellik C4.5 modeline ait karar ağacı ise Şekil 7'de görülmektedir. 12 özellik kullanılarak oluşturulan Rastgele Orman modeli yoğun bakım tedavisine ihtiyaç duyan hastaları 0,954 F1-Skor ile sınıflandırırken veri sayısının az olmasından dolayı yoğun bakım tedavisine ihtiyaç duymayan hastaları 0,735 F1-Skoru ile ortalama bir sınıflandırma performansıyla tahmin etmiştir.

Tablo 13 12 özellik kullanılarak ve SMOTE uygulandıktan sonra 12 özellik kullanılarak oluşturulan model sonuçlarının doğruluk ve F1-skor karşılaştırılmasını temsil etmektedir. %95,26 değeri ile SMOTE tekniği uygulandıktan sonra genel doğruluk değeri açısından yaklaşık %3'lük bir artış görülmüştür. SMOTE kullanılmadan önce 0,74 F1-



Şekil 6. Analiz – 2 SMOTE uygulanmayan ve uygulanan modellerin karmaşıklık matrisi
(Analysis - 2 Confusion matrix of non-SMOTE and applied models)

Tablo 10. Analiz – 3 seçilen özellikler (Analysis – 3 selected features)

Seçilen Özellikler						
Cinsiyet	HCT	HGB	Kütle CK-MB	MDW	Methemoglobin	MON#
MON%	PO2	RBC	Troponin T-hs	WHO giriş		

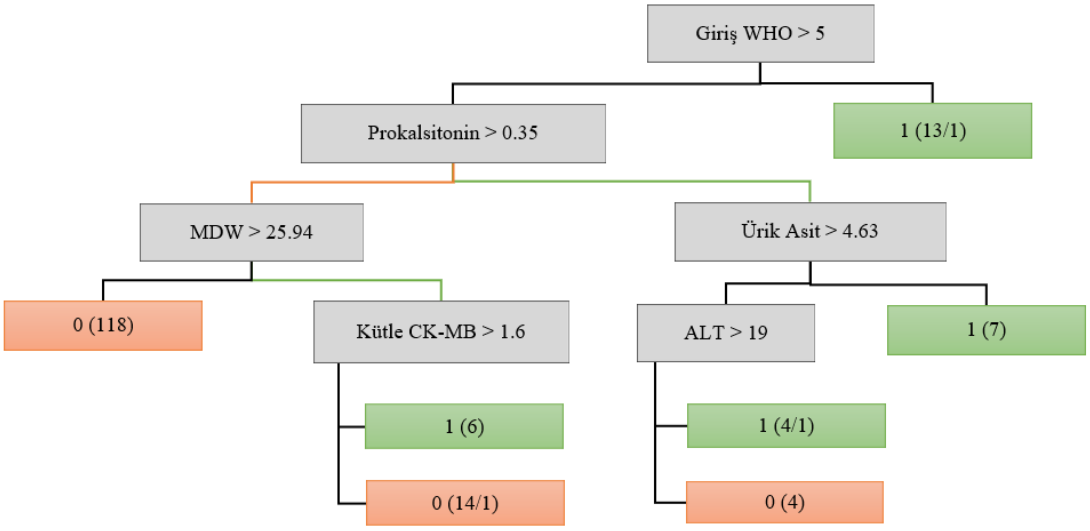
Tablo 11. Analiz – 3 Tüm özellikler ve seçilen özellikler ile oluşturulan modellerin doğruluk ve model parametre değerleri
(Analysis – 3 performance and parameters of models created with all features and selected features)

Özellik Sayısı	Algoritma	Doğruluk	Parametre
73	KNN	%89,16	K: 5
	Bagging	%89,16	Sınıflandırıcı: RepTree
	Rastgele Orman	%89,16	Ağaç sayısı: 60
	Karar Ağacı	*%92,77*	(Minimum obje sayısı: 4, Güven faktörü: 0,1)
12	KNN	%88,55	K: 8
	Bagging	%90,96	Sınıflandırıcı: J48
	Rastgele Orman	*%92,17*	Ağaç sayısı: 100
	Karar Ağacı	%91,57	(Minimum obje sayısı: 4, Güven faktörü: 0,2)

Tablo 12. Analiz – 3 Tüm özellikler ve seçilen özellikler ile oluşturulan modellerin performansları ve
(Analysis – 3 performance of models created with all features and selected features)

Model Adı	Doğruluk	Yoğun bakıma ihtiyaç	F1-Skor
73 özellik C4.5 Karar Ağacı	*%92,77*	VAR	*0,786*
		YOK	*0,957*
12 özellik Rastgele Orman	%92,17	VAR	0,735
		YOK	0,954

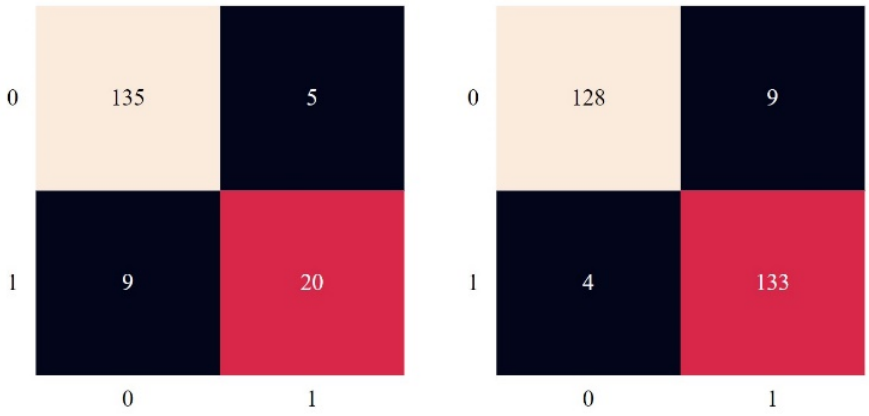
Skoruyla tahmin edilen oksijen tedavisine ihtiyaç duyan hastalar SMOTE uygulandıktan sonra 0,95 F1-Skor değerine yükselmiştir. Şekil 8 sırayla 12 özellikli SMOTE uygulanmayan ve uygulanan modellerin karmaşıklık matrislerini temsil etmektedir. Karmaşıklık matrislerine göre SMOTE uygulanmadan önce "1" etiketli oksijen tedavisine ihtiyaç duyan sekiz hasta yanlış sınıflandırılırken SMOTE uygulandıktan sonra bu değer dörde düşmüştür. Yanlış sınıflandırılan "1" etiketli hastalardan 3 tanesi yapay bir veriler iken ve kalan veri ise gerçek bir hasta kaydına aittir.



Şekil 7. Analiz – 3 73 özellikli Karar Ağacı Modeli (Analysis – 3 Decision Tree Model with 73 features)

Tablo 13. Analiz – 3 SMOTE uygulanan ve uygulanmayan modellerin performansları (Analysis – 3 performance of models with and without SMOTE)

Model Adı	Doğruluk	Yoğun bakıma ihtiyaç	F1-Skor
12 özellik Rastgele Orman	%92,17	VAR	0,74
		YOK	0,95
SMOTE - 12 özellik Rastgele Orman	*%95,26*	VAR	0,95
		YOK	0,95



Şekil 8. Analiz – 3 SMOTE uygulanmayan ve uygulanan modellerin karmaşıklık matrisi (Analysis - 3 Confusion matrix of non-SMOTE and applied models)

Tablo 14. Özellik Seçme işleminin sınıflandırma performasına etkisi (Effect of Feature Selection on classification performance)

Analiz	Özellik Sayısı	Hedef Değer	Algoritma	Doğruluk
Analiz - 1	72	WHO giriş	Rastgele Orman	89,44
Analiz - 1	16	WHO giriş	Rastgele Orman	91,67
Analiz - 2	73	WHO en kötü	C4.5 Karar Ağacı	88,39
Analiz - 2	18	WHO en kötü	Rastgele Orman	91,96
Analiz - 3	73	WHO en kötü	C4.5 Karar Ağacı	92,77
Analiz - 3	12	WHO en kötü	Rastgele Orman	92,17

Tablo 15. SMOTE veri çoğaltmanın sınıflandırma performasına etkisi (Effect of SMOTE on classification performance)

Analiz Numarası	SMOTE	Doğruluk (%)	Sınıf Etiketi	F1- Skor (%)
1	YOK	91,67	Oksijen tedavisi ihtiyacı var	89
			Oksijen tedavisi ihtiyacı yok	93
	VAR	96,64	Oksijen tedavisi ihtiyacı var	95
			Oksijen tedavisi ihtiyacı yok	95
2	YOK	91,96	Oksijen tedavisi ihtiyacı var	74
			Oksijen tedavisi ihtiyacı yok	95
	VAR	97,8	Oksijen tedavisi ihtiyacı var	98
			Oksijen tedavisi ihtiyacı yok	98
3	YOK	92,17	Yoğun bakım ihtiyacı var	74
			Yoğun bakım ihtiyacı yok	95
	VAR	95,26	Yoğun bakım ihtiyacı var	95
			Yoğun bakım ihtiyacı yok	95

Tablo 14 her bir analiz için tüm özellikler ve seçilen özellikler kullanılarak oluşturulan modellerin doğruluk değerlerini temsil etmektedir. Sonuçlar 3 analiz içinde daha az sayıda özellik kullanılarak yüksek doğruluk değerlerine ulaşabildiğini göstermektedir.

Tablo 15 ise SMOTE algoritması kullanılmadan önce ve sonra elde edilen performans metrik değerlerini temsil etmektedir. Sonuçlar üç analiz için de genel sınıflandırma doğruluğunda SMOTE uygulandıktan yükselme olduğunu göstermektedir.

4. Tartışma (Discussions)

Çalışma kapsamında yapılan analizlerde seçilen özellikler ile literatürdeki çalışmalar sonucunda seçilen özellikler karşılaştırıldığında CRP, Albumin, Ferritin, D-dimer, Kalsiyum, NEU, WBC, Direkt bilirubin ve LDH niteliklerinin ortak olarak seçildiği görülmüştür [38-40].

Analiz-1 ve Analiz-2'de hedeflenen solunum ihtiyacı ile en ilişkili özellikler arasında seçilen CRP ve LDH Olmedo ve arkadaşları [41] tarafından yapılan çalışmada da en ilişkili olarak seçilen 4 özellik içerisinde yer almaktadır. Ayrıca bu çalışma ile ortak olarak Arvind vd. [13] CRP, D-dimer ve lökosit/lenfosit sayısı (WBC) laboratuvar değişkenlerinin hastanede yatan COVID-19 hastalarında entübasyon riskini tahmin etmede daha yüksek bir öneme sahip olduğunu da göstermiştir.

Dengesiz veri kümeleri uygun bir şekilde dengelenmediği sürece daha düşük doğruluğa sahip modellerin oluşmasına yol açmaktadır. SMOTE algoritması dengesiz veri kümesiyle çalışmak için en yaygın kullanılan yüksek örnekleme (oversampling) tekniklerinden biridir. SMOTE tekniğinin kullanılması bu çalışmadaki sınıflama analiz sonuçlarına göre sınıflama başarısını artırmaktadır. Aljameel vd. [42], Sowjanya ve Mrudula. [43]' nın çalışmalarında olduğu gibi SMOTE yönteminin sağlık verisi üzerinde kullanılmasının sınıflandırma doğruluğunda pozitif bir artış sağladığı çalışma sonuçlarından anlaşılmaktadır.

Sonuçlar incelendiğinde her üç analiz içinde genel sınıflandırma doğruluğunda SMOTE algoritması uygulandıktan sonra yükselme olduğu görülmektedir. Ayrıca azınlık olan sınıfların F1-Skor değerleri karşılaştırıldığında Analiz – 1 için %89'dan %95'e olacak şekilde %6'lik bir artış, Analiz – 2 için %74'ten % 98'ye olacak şekilde %24'lik bir artış ve Analiz – 3 için %74'ten %95'e olacak şekilde %21'lik bir artış gözlendiği sonucuna varılmaktadır. Bu da SMOTE algoritması uygulandıktan sonra azınlık olan sınıfın daha iyi bir sınıflandırma performansı ile tahmin edilebildiğini göstermektedir.

5. Sonuçlar (Conclusions)

Bu çalışmada gerçekleştiren Analiz-1'de hastanın hastaneye yatış aşamasındaki oksijen ihtiyacının tahmin edilmesi hedeflenmiştir. Laboratuvar ve demografik olmak üzere toplam 72 özellik ve WHO giriş hedef değeri kullanılarak modeller oluşturulmuştur. 72 özelliğin hepsi kullanılarak en yüksek doğruluk değerine %89,44 ile Rastgele Orman modeli ulaşırken özellik seçme algoritması ile bulunan 16 özellik kullanılarak oluşturulan Rastgele Orman modeli %91,67 doğruluk değerine ulaşmıştır. Sonuçlar 16 özellik kullanılarak daha yüksek bir sınıflandırma doğruluğuna ulaşıldığını göstermektedir.

Hastanın hastanede yatış süresi boyunca oksijen ihtiyacının tahmin edilmesi için gerçekleştirilen Analiz-2 çalışmasında ise laboratuvar, demografik ve WHO giriş dahil olmak üzere toplam 73 özellik ve WHO en kötü hedef değeri kullanılarak modeller oluşturulmuştur. 73 özelliğin hepsi kullanılarak en yüksek doğruluk değerine %88,39 ile C4.5 Karar Ağacı modeli ulaşırken özellik seçme algoritması ile bulunan 18 özellik kullanılarak oluşturulan Rastgele Orman modeli %91,96 doğruluk değerine ulaşmıştır. Sonuçlar 18 özellik kullanılarak daha yüksek bir sınıflandırma doğruluğuna ulaşıldığını göstermektedir.

Analiz-3'de hastanın hastanede yatış süresi boyunca yoğun bakım ihtiyacının tahmin edilmesi hedeflenmiştir. Laboratuvar, demografik ve WHO giriş dahil olmak üzere toplam 73 özellik ve WHO en kötü hedef değeri kullanılarak modeller oluşturulmuştur. 73 özelliğin hepsi kullanılarak en yüksek doğruluk değerine %92,77 ile C4.5 Karar Ağacı modeli ulaşırken özellik seçme algoritması ile bulunan 12 özellik kullanılarak oluşturulan Rastgele Orman modeli %91,96 doğruluk değerine ulaşmıştır. Sonuçlar 12 özellik kullanılarak yüksek bir sınıflandırma doğruluğuna ulaşıldığını göstermektedir.

MDW ve Troponin T-hs her üç analizde de ortak olarak seçilen en ilişkili özelliklerdir. Ayrıca WHO ölçüt değeri pandemi başında ve öncesindeki süreçte çoğunlukla hastalık şiddeti ciddi olan hastaları tespit etmek amaçlı kullanıldığından ve bu ölçüt elde edilirken kullanılan parametreler çoğunlukla yoğun bakım durumunu belirleyen parametreler olduğundan dolayı Analiz-3 için seçilen ilişkili özellikler arasında yer alırken Analiz-2 kapsamında seçilen özellikler içerisinde yer almamaktadır.

COVID-19 enfeksiyonunun sonucunu tahmin etmek için hangi biyokimyasal veya hematolojik belirteçlerin kullanılacağının belirlenmesi önemli bir araştırma konusudur. Çalışmamız COVID-19 ile ilgili önemli hasta laboratuvar niteliklerini belirlemesi açısından bu alandaki çalışmalara katkı sağlayan sonuçlara sahip bir araştırmadır. Çalışmamızın sonuçlarının sağlık hizmetleri kaynaklarının kullanımının iyileştirilmesine ve COVID-19 riskini değerlendiren tahmine dayalı modellerin kullanımına katkısı bulunmaktadır.

Çalışma kapsamında elde edilen bu sonuçlar ile yeterli sayıda bulunmayan yoğun bakım ve oksijen gibi tıbbi kaynaklar için uygun bir kullanım stratejisi ve etkili bir risk sınıflandırma, triyaj protokolü oluşturulmasına yardımcı olunması hedeflenmiştir. Bu çalışmada laboratuvar testleri sonuçları kullanılarak analiz işlemi yapıldığından dolayı tıbbi kaynak sıkıntısı çeken bölgeler için faydalı bir çalışma olabileceği düşünülmektedir.

Teşekkür (Acknowledgement)

Çalışmadaki veri toplama ve araştırma süreci tasarımındaki destekleri için Marmara Üniversitesi Hipertansiyon ve Ateroskleroz Eğitim, Uygulama ve Araştırma Merkezi (HİPAM) müdürü Sayın Prof. Dr. Ali Serdar Fak ve araştırmacılarına teşekkürlerimizi sunarız.

Kaynaklar (References)

1. Guarner J., Three Emerging Coronaviruses in Two Decades, Am J Clin Pathol, 153 (4), 420-421, 2020.
2. Cheruku S.R., Barina A., Kershaw C.D., Goff K., Reisch J., Hynan L.S., Ahmed F., Armaignac D.L., Patel L. Belden K.A., Palliative care consultation and end-of-life outcomes in hospitalized COVID-19 patients, Resuscitation, 170, 230-237, 2022.
3. Organization P.A.H. WHO characterizes COVID-19 as a pandemic. https://www3.paho.org/hq/index.php?option=com_content&view=article&id=15756:who-characterizes-covid-19-as-a-pandemic&Itemid=1926&lang=en. Yayın tarihi 2022. Erişim tarihi Temmuz 6, 2023.
4. Organization W.H. WHO coronavirus (COVID-19) emergency dashboard. https://covid19.who.int. Yayın tarihi 2021. Erişim tarihi Temmuz 6, 2023.
5. Fontanarosa P.B. Bauchner H., COVID-19—looking beyond tomorrow for health care and society, Jama, 323 (19), 1907-1908, 2020.
6. Huang C., Wang Y., Li X., Ren L., Zhao J., Hu Y., Zhang L., Fan G., Xu J., Gu X., Cheng Z., Yu T., Xia J., Wei Y., Wu W., Xie X., Yin W., Li H., Liu M., Xiao Y., Gao H., Guo L., Xie J., Wang G., Jiang R., Gao Z., Jin Q., Wang J. Cao B., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet, 395 (10223), 497-506, 2020.
7. Wang D., Hu B., Hu C., Zhu F., Liu X., Zhang J., Wang B., Xiang H., Cheng Z., Xiong Y., Zhao Y., Li Y., Wang X. Peng Z., Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China, Jama, 323 (11), 1061-1069, 2020.
8. Çilgin C., Gökçen H. Gökşen Y., Sentiment analysis of public sensitivity to COVID-19 vaccines on twitter by majority voting classifier-based machine learning, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (2), 1093-1104, 2022.
9. Sönmez N., Terim Cavka B., Recommendations for the transformation of patient rooms into isolated patient rooms in the process of the COVID-19 pandemic, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (1), 175-188, 2022.
10. Banerjee A., Ray S., Vorselaars B., Kitson J., Mamalakis M., Weeks S., Baker M. Mackenzie L.S., Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population, Int Immunopharmacol, 86, 106705, 2020.
11. Mondal M.R.H., Bharati S., Podder P. Podder P., Data analytics for novel coronavirus disease, Informatics in Medicine Unlocked, 20, 100374, 2020.
12. Akarsu E., Classification of Coronavirus Disease with Artificial Intelligence and Machine Learning, Avrupa Bilim ve Teknoloji Dergisi, (36), 6-9, 2022.
13. Arvind V., Kim J.S., Cho B.H., Geng E. Cho S.K., Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19, J Crit Care, 62, 25-30, 2021.
14. Burdick H., Lam C., Mataraso S., Siefkas A., Braden G., Dellinger R.P., McCoy A., Vincent J.L., Green-Saxena A., Barnes G., Hoffman J., Calvert J., Pellegrini E. Das R., Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial, Comput Biol Med, 124, 103949, 2020.
15. Di Castelnuovo A., Bonaccio M., Costanzo S., Gialluisi A., Antinori A., Berselli N., Blandi L., Bruno R., Cauda R., Guaraldi G., My I., Menicanti L., Parruti G., Patti G., Perlini S., Santilli F., Signorelli C., Stefanini G.G., Vergori A., Abdeddaim A., Ageno W., Agodi A., Agostoni P., Aiello L., Al Moghazi S., Aucella F., Barbieri G., Bartoloni A., Bologna C., Bonfanti P., Brancati S., Cacciatore F., Caiano L., Cannata F., Carrozzi L., Cascio A., Cingolani A., Cipollone F., Colomba C., Crisetti A., Crosta F., Danzi G.B., D'Ardes D., de Gaetano Donati K., Di Gennaro F., Di Palma G., Di Tano G., Fantoni M., Filippini T., Fioretto P., Fusco F.M., Gentile I., Grisafi L., Guarnieri G., Landi F., Larizza G., Leone A., Maccagni G., Maccarella S., Mapelli M., Maragna R., Marcucci R., Maresca G., Marotta C., Marra L., Mastroianni F., Mengozzi A., Menichetti F., Milic J., Murri R., Montineri A., Mussinelli R., Mussini C., Musso M., Odone A., Olivieri M., Pasi E., Petri F., Pinchera B., Pivato C.A., Pizzi R., Poletti V., Raffaelli F., Ravaglia C., Righetti G., Rognoni A., Rossato M., Rossi M., Sabena A., Salinaro F., Sangiovanni V., Sanrocco C., Scarafino A., Scorzolini L., Sgariglia R., Simeone P.G., Spinoni E., Torti C., Trecarichi E.M., Vezzani F., Veronesi G., Vettor R., Vianello A., Vinceti M., De Caterina R. Iacoviello L., Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study, Nutr Metab Cardiovasc Dis, 30 (11), 1899-1913, 2020.
16. Huyut M.T., Automatic Detection of Severely and Mildly Infected COVID-19 Patients with Supervised Machine Learning Models, IRBM, 44 (1), 100725, 2023.
17. Gözde Ş., Demirel E., Selen A. Aladağ Z., Evaluation of effective risk factors in COVID-19 mortality rate with DEMATEL method, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (4), 2151-2166, 2021.
18. Huyut M.T. Üstündağ H., Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study, Med Gas Res, 12 (2), 60-66, 2022.
19. Huyut M.T. Huyut Z., Forecasting of Oxidant/Antioxidant levels of COVID-19 patients by using Expert models with biomarkers used in the Diagnosis/Prognosis of COVID-19, Int Immunopharmacol, 100, 108127, 2021.
20. Cabitza F., Campagner A., Ferrari D., Resta C.D., Ceriotti D., Sabetta E., Colombini A., Vecchi E.D., Banfi G., Locatelli M. Carobene A., Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests, Clinical Chemistry and Laboratory Medicine (CCLM), 59 (2), 421-431, 2021.
21. Marshall J.C., Murthy S., Diaz J., Adhikari N., Angus D.C., Arabi Y.M., Baillie K., Bauer M., Berry S. Blackwood B., A minimal common outcome measure set for COVID-19 clinical research, The Lancet Infectious Diseases, 20 (8), e192-e197, 2020.
22. Napoleon D. Pavalakodi S., A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set, International Journal of Computer Applications, 13 (7), 41-46, 2011.
23. Wei J., Research on data preprocessing in supermarket customers data mining, 2010 2nd International Conference on Information Engineering and Computer Science, 1-4, 2010.
24. Kaiser J., Dealing with Missing Values in Data, Journal of Systems Integration, 5 (1), 2014.
25. Grossman R.L., Kamath C., Kegelmeyer P., Kumar V. Namburu R., Data mining for scientific and engineering applications, Cilt 2, Springer Science & Business Media, 2013.
26. Padmaja D.L. Vishnuvardhan B., Comparative study of feature subset selection methods for dimensionality reduction on scientific data, 2016 IEEE 6th International Conference on Advanced Computing (IACC), 31-34, 2016.
27. Read B.J., Data mining and science? Knowledge discovery in science as opposed to business, 1999.
28. Pereira R.B., Plastino A., Zadrozny B. Merschmann L.H., Categorizing feature selection methods for multi-label classification, Artificial Intelligence Review, 49 (1), 57-78, 2018.
29. Chawla N.V., Bowyer K.W., Hall L.O. Kegelmeyer W.P., SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321-357, 2002.
30. Aydilek İ.B., Yazılım hata tahmininde kullanılan metriklerin karar ağaçlarındaki bilgi kazançlarının incelenmesi ve iyileştirilmesi, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24 (5), 906-914, 2018.

31. Bhargava N., Sharma S., Purohit R. Rathore P.S., Prediction of recurrence cancer using J48 algorithm, 2017 2nd International Conference on Communication and Electronics Systems (ICCES), 386-390, 2017.

32. Yadav S. Shukla S., Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification, 2016 IEEE 6th International conference on advanced computing (IACC), 78-83, 2016.

33. Landron C., Development of an amorphous film microanalysis method, Thin Solid Films, 84 (2), 143-144, 1981.

34. Sun Y., Zhang H., Zhao T., Zou Z., Shen B. Yang L., A new convolutional neural network with random forest method for hydrogen sensor fault diagnosis, IEEE Access, 8, 85421-85430, 2020.

35. Breiman L., Bagging predictors, Machine learning, 24 (2), 123-140, 1996.

36. Quinlan J.R., Simplifying decision trees, International journal of man-machine studies, 27 (3), 221-234, 1987.

37. Xing B., Zhang H., Zhang K., Zhang L., Wu X., Shi X., Yu S. Zhang S., Exploiting EEG signals and audiovisual feature fusion for video emotion recognition, IEEE Access, 7, 59844-59861, 2019.

38. Alballa N. Al-Turaiki I., Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review, Informatics in medicine unlocked, 24, 100564, 2021.

39. Huang I., Pranata R., Lim M.A., Oehadian A. Alisjahbana B., C-reactive protein, procalcitonin, D-dimer, and ferritin in severe coronavirus disease-2019: a meta-analysis, Therapeutic advances in respiratory disease, 14, 1753466620937175, 2020.

40. Li Q., Cao Y., Chen L., Wu D., Yu J., Wang H., He W., Chen L., Dong F. Chen W., Hematological features of persons with COVID-19, Leukemia, 34 (8), 2163-2172, 2020.

41. Domínguez-Olmedo J.L., Gragera-Martínez Á., Mata J. Pachón Álvarez V., Machine learning applied to clinical laboratory data in Spain for COVID-19 outcome prediction: model development and validation, Journal of medical Internet research, 23 (4), e26211, 2021.

42. Aljameel S.S., Khan I.U., Aslam N., Aljabri M. Alsulmi E.S., Machine learning-based model to predict the disease severity and outcome in COVID-19 patients, Scientific programming, 2021, 1-10, 2021.

43. Sowjanya A.M. Mrudula O., Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms, Applied Nanoscience, 13 (3), 1829-1840, 2023.